ECE 6882 Usama Yousuf

Homework 3 Date:

Problem 1: Policy Iteration

Initialization: $s_k \in \{s^1 = 0, s^2 = 1\}$

$$V_0(s^1) = V_0(s^2) = 0$$

$$\pi_0(s^1) = \pi_0(s^2) = a^1$$

$$\gamma = 0.9, \theta = 0.85$$

Evaluation:

Iter 1: $\Delta = 0$

For s_1:

$$v = V_0(s^1) = 0, \quad \pi_0(s^1) = a^1$$

$$\Rightarrow V_1(s^1) = p(s^1, r \mid s^1, \pi_0(s^1)) [r + \gamma V_0(s^1)] + p(s^2, r \mid s^1, \pi_0(s^1)) [r + \gamma V_0(s^1)]$$

$$\Rightarrow V_1(s^1) = p(s^1 \mid s^1, a^1) [r + \gamma V_0(s^1)] + p(s^2 \mid s^1, a^1) [r + \gamma V_0(s^1)]$$

$$= 1 \cdot [0 + 0.9 \times 0] + 0$$

$$\Rightarrow \boxed{V_1(s^1) = 0}$$

$$\Delta = \max(0, |0-0|) = 0$$

For s_2:

$$v = V_0(s^2) = 0$$

$$V_1(s^2) = p(s^1 \mid s^2, a^1) [r + \gamma V_0(s^1)] + p(s^2 \mid s^2, a^1) [r + \gamma V_0(s^2)]$$

$$= 0 + 1 [1 + 0.9 \times 0] = 1$$

$$\boxed{V_1(s^2) = 1}$$

$$\Delta = \max(0, |0-1|) = 1$$

$\Delta < 0.85$ is false,

so we continue.

Date:

Iteration 2:

$\Delta = 0, V_1(s^1) = 0, V_1(s^2) = 1$

For s^1:

$V_2(s^1) = p(s^2|s^1, a^1)[1 + 0.9 \times 1] +$
$p(s^1|s^1, a^1)[0 + 0.9 \times 0]$

$V_2(s^1) = 0 + 0 = 0, \quad \Delta = 0$

For s^2:

$V_2(s^2) = p(s^2|s^2, a^1)[1 + 0.9 \times 1] +$
$p(s^1|s^2, a^1)[0 + 0.9 \times 0]$

$V_2(s^2) = 1[1 + 0.9] + 0 = 1.9$

$\Delta = \max(0, |1 - 1.9|) = 0.9$

$\Delta < 0.85$ is false, continue.

Iteration 3:

$\Delta = 0, V_2(s^1) = 0; V_2(s^2) = 1.9$

For s^1:

$V_2(s^1) = p(s^1|s^1, a^1)[0 + 0.9 \times 0] +$
$p(s^2|s^1, a^1)[1 + 0.9 \times 1.9]$

$\boxed{V_2(s^1) = 0} \quad \Delta = 0$

For s^2:

$V_2(s^2) = p(s^1|s^2, a^1)[0 + 0.9 \times 0] +$
$p(s^2|s^2, a^1)[1 + 0.9 \times 1.9]$

$= 0 + 1[1 + 1.71]$

$\boxed{V_2(s^2) = 2.71}$

$\Delta = \max(0, |2.71 - 1.9|) = 0.81$

$\Delta < 0.85$ is true!

Date:

Policy Improvement:

stable = True

For s^1:

old_action = a^1

$$\pi_1(s^1) = \underset{a}{\text{argmax}} \begin{cases} p(s^1|s^1,a^1)[0+0.9\times 0] + p(s^2|s^1,a^1)[1+0.9\times 2.71], \\ p(s^1|s^1,a^2)[0+0.9\times 0] + p(s^2|s^1,a^2)[1+0.9\times 2.71] \end{cases}$$

$$= \underset{a}{\text{argmax}} \{ \underbrace{0+0}_{a^1}, \underbrace{0+1(3.43)}_{a^2} \}$$

$$\Rightarrow \boxed{\pi_1(s^1) = a^2}$$

For s^2:

$$\pi_1(s^2) = \underset{a}{\text{argmax}} \begin{cases} p(s^1|s^2,a^1)[0+0.9\times 0] + p(s^2|s^2,a^1)[1+0.9\times 2.71], \\ p(s^1|s^2,a^2)[0+0.9\times 0] + p(s^2|s^2,a^2)[1+0.9\times 2.71] \end{cases}$$

$$= \underset{a}{\text{argmax}} \begin{cases} 0+1[1+0.9\times 2.71] & \leftarrow a_1 \\ 0+0 & \leftarrow a_2 \end{cases}$$

$$= \underset{a}{\text{argmax}} \{ 3.43, 0 \} = a_1$$

$$\Rightarrow \boxed{\pi_1(s^2) = a^1}$$

Thus,

$$\pi_1(s^1) = a^2, \quad \pi_1(s^2) = a^1.$$

Date:

Problem 2: Value Iteration.

a) $V = \begin{bmatrix} V_0(s^1) \\ V_0(s^2) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\gamma = 0.9$

Compute V_1, V_2, V_3.

→ For V_1:

$$\Rightarrow V_1(s^1) = \max_a \{ p(s^1|s^1,a^1)[0+0.9\times 0] + p(s^2|s^1,a^1)[1+0.9\times 0], \; p(s^1|s^1,a^2)[0+0.9\times 0] + p(s^2|s^1,a^2)[1+0.9\times 0] \}$$

$$V_1(s^1) = \max[0+0, 0+1]$$

$$\boxed{V_1(s^1) = 1}$$

$$\Rightarrow V_1(s^2) = \max_a \{ p(s^1|s^2,a^1)[0+0.9\times 0] + p(s^2|s^2,a^1)[1+0.9\times 0], \; p(s^1|s^2,a^2)[0+0.9\times 0] + p(s^2|s^2,a^2)[1+0.9\times 0] \}$$

$$= \max_a \{ 0 + 1\times 1, 0 \}$$

$$\Rightarrow \boxed{V_1(s^2) = 1}$$

$$\Rightarrow V_1 = \begin{bmatrix} V_1(s^1) \\ V_1(s^2) \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

Date:

→ For V_2:

$$\Rightarrow V_2(s^1) = \max_a \{ \cancel{p} \; 1.(0+0.9)+0.,\ 1.(1+0.9\times 1) \}$$

$$= \max_a \{ 0.9, 1.9 \} = 1.9$$

$$\Rightarrow V_2(s^2) = \max_a \{ 1.(1+0.9\times 1), 1(0+0.9\times \cancel{1}) \}$$

$$= \max_a \{ 1.9, 0.9 \} = 1.9$$

$$\Rightarrow V_2 = \begin{bmatrix} V_2(s^1) \\ V_2(s^2) \end{bmatrix} = \begin{bmatrix} 1.9 \\ 1.9 \end{bmatrix}$$

For V_3:

$$\Rightarrow V_3(s^1) = \max_a \{ 1(0+0.9\times 1.9), 1(1+0.9\times 1.9) \}$$

$$= \max_a \{ 1.71, 2.71 \}$$

$$\boxed{V_3(s^1) = 2.71}$$

$$\Rightarrow V_3(s^2) = \max_a \{ 1(1+0.9\times 1.9),\ 1(0+0.9\times 1.9) \}$$

$$= \max \{ 2.71, 1.71 \}$$

$$\Rightarrow \boxed{V_3(s^2) = 2.71}$$

Thus, $V_3 = \begin{bmatrix} V_3(s^1) \\ V_3(s^2) \end{bmatrix} = \begin{bmatrix} 2.71 \\ 2.71 \end{bmatrix}$

Date:

$(s^1=0, s^2=1)$

b) Compute $\pi(s=0)$ & $\pi(s=1)$.

$$V_3 = \begin{bmatrix} 2.71 \\ 2.71 \end{bmatrix}$$

$$\pi(s=0) = \underset{a}{\operatorname{argmax}} \begin{Bmatrix} p(s^1|s^1,a^1)(0+0.9\times2.71), \\ p(s^2|s^1,a^2)(1+0.9\times2.71) \end{Bmatrix}$$

$$= \underset{a}{\operatorname{argmax}} \{ \underset{a^1}{\underbrace{1(2.43)}}, \underset{a^2}{\underbrace{3.43}} \}$$

$$\Rightarrow \boxed{\pi(s=0) = a^2}$$

$$\pi(s=1) = \underset{a}{\operatorname{argmax}} \begin{Bmatrix} p(s^2|s^2,a^1)(1+0.9\times2.71), \\ p(s^1|s^2,a^2)(0+0.9\times2.71) \end{Bmatrix}$$

$$= \underset{a}{\operatorname{argmax}} \{ \underset{a_1}{\underbrace{3.43}}, \underset{a^2}{\underbrace{2.43}} \}$$

$$\boxed{\pi(s=1) = a^1}$$

Thus, $\pi = \begin{bmatrix} \pi(s=0) \\ \pi(s=1) \end{bmatrix} = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$

Date:

Problem 3: Non-deterministic Policy Iteration

$$\gamma = 0.9 \quad , \quad \pi(s) = \begin{bmatrix} \pi(A) \\ \pi(B) \end{bmatrix} = \begin{bmatrix} a^2 \\ a^2 \end{bmatrix}$$

$$M(a^1) = \begin{array}{c|cc} & A & B \\ \hline A & 1 & 0 \\ B & 0.8 & 0.2 \end{array}$$

$$M(a^2) = \begin{array}{c|cc} & A & B \\ \hline A & 0.2 & 0.8 \\ B & 0.8 & 0.2 \end{array}$$

~~Assume~~ $\cancel{V^{\pi}(s) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = \begin{bmatrix} V^{\pi}(A) \\ V^{\pi}(B) \end{bmatrix}}$

Policy Evaluation:

$$V(A) = P(A|A,a^2) \times (R(A,a^2,A) + \gamma V(A)) + P(B|A,a^2) \times (R(A,a^2,B) + \gamma V(B))$$

$$\Rightarrow V(A) = 0.2(-10 + 0.9 \times V(A)) + 0.8(-5 + 0.9 \times V(B))$$

$$\Rightarrow V(A) = -2 + 0.18V(A) - 4 + 0.72V(B)$$

$$\Rightarrow 0.82V(A) - 0.72V(B) + 6 = 0 \quad - Eq\ 1$$

$$V(B) = P(A|B,a^2) \times (R(B,a^2,A) + \gamma V(A)) + P(B|B,a^2) \times (R(B,a^2,B) + \gamma V(B))$$

$$V(B) = 0.8 \times (20 + 0.9V(A)) + 0.2 \times (10 + 0.9V(B))$$

$$V(B) = 16 + 0.72V(A) + 2 + 0.18V(B)$$

$$\Rightarrow 0.82V(B) - 0.72V(A) - 18 = 0 \quad - Eq.\ 2$$

Date:

Solving Eq 1 & 2 simultaneously,

$$1 \Rightarrow 0.82V(A) = 0.72V(B) - 6$$

$$\Rightarrow \boxed{V(A) = \frac{0.72V(B)-6}{0.82}} \quad - 3$$

Substitute in 2,

$$2 \Rightarrow 0.82V(B) - 0.72\left(\frac{0.72V(B)-6}{0.82}\right) - 18 = 0$$

$$\Rightarrow 0.82V(B) - \left(\frac{0.5184V(B) - 4.32}{0.82}\right) - 18 = 0$$

$$\Rightarrow 0.6724V(B) - 0.5184V(B) + 4.32 - 14.76 = 0$$

$$\Rightarrow \cancel{0.5184}$$

$$\Rightarrow 0.154V(B) - 10.44 = 0$$

$$\Rightarrow V(B) = 10.44/0.154$$

$$\Rightarrow \boxed{V(B) = 67.792}$$

Substitute back in 3,

$$3 \Rightarrow V(A) = \frac{0.72 \times 67.792 - 6}{0.82}$$

$$\boxed{V(A) = 52.2}$$

Thus, $V^{\pi}(s) = \begin{bmatrix} V(A) \\ V(B) \end{bmatrix} = \begin{bmatrix} 52.2 \\ 67.792 \end{bmatrix}$

Date:

Policy Improvement:

~~stable = True.~~

~~For A:~~

~~old action~~

Since we have the optimal state values $V^{\pi}(s)$ already, we can directly determine optimal policy as follows:

$$\pi(A) = \underset{a}{\text{argmax}} \begin{cases} P(A|A,a^1)\{R(A,a^1,A)+\gamma.V^{\pi}(A)\}+ \\ P(B|A,a^1)\{R(A,a^1,B)+\gamma.V^{\pi}(B)\}, \\ P(A|A,a^2)\{R(A,a^2,A)+\gamma.V^{\pi}(A)\}+ \\ P(B|A,a^2)\{R(A,a^2,B)+\gamma.V^{\pi}(B)\} \end{cases}$$

$$\pi(A) = \underset{a}{\text{argmax}} \begin{cases} 1.(10+0.9\times52.2)+0, & \leftarrow a_1 \\ 0.2(-10+0.9\times52.2) + \\ 0.8(-5+0.9\times67.792) & \} a_2 \end{cases}$$

$$= \underset{a}{\text{argmax}} \{56.98, 52.2\}$$

$$\Rightarrow \boxed{\pi(A) = a^1}$$

$$\pi(B) = \underset{a}{\text{argmax}} \begin{cases} P(A|B,a^1)\{R(B,a^1,A)+\gamma V^{\pi}(A)\}+ \\ P(B|B,a^1)\{R(B,a^1,B)+\gamma V^{\pi}(B)\}, \\ P(A|B,a^2)\{R(B,a^2,A)+\gamma V^{\pi}(A)\}+ \\ P(B|B,a^2)\{R(B,a^2,B)+\gamma V^{\pi}(B)\} \end{cases}$$

Date:

$$\pi(B) = \underset{a}{\operatorname{argmax}} \{0.8(40+0.9\times52.2) + 0.2(10+0.9\times67.792),\ 0.8(20+0.9\times52.2) + 0.2(10+0.9\times67.792)\}$$

$$= \underset{a}{\operatorname{argmax}} \{83.78, 67.786\}$$

$$\Rightarrow \boxed{\pi(B) = a'}$$

Thus,

Optimal policy:

$$\pi^*(s) = \begin{bmatrix} a' \\ a' \end{bmatrix} = \begin{bmatrix} \pi^*(s=A) \\ \pi^*(s=B) \end{bmatrix}$$

State Values:

$$V^{\pi}(s) = \begin{bmatrix} 52.2 \\ 67.792 \end{bmatrix} = \begin{bmatrix} V^{\pi}(s=A) \\ V^{\pi}(s=B) \end{bmatrix}$$

× ———— × ———— ×